# Face Recognition and Tagging

Valluri . Vinay Teja

Computer Science and Engineering

Lovely professional University Jalandhar,

India.

Desina. Sai Vivek

Computer Science and Engineering

Lovely professional University Jalandhar,

India.

**Abstract:** Face recognition is currently becoming popular to be applied in various ways, especially in security systems. Various methods of face recognition have been proposed in researches and increased accuracy is the main goal in the development of face recognition methods. FaceNet is one of the new method in face recognition technology. FaceNet is a deep neural network used for extracting features from an image of a person's face. Ideally, embeddings of similar faces are also similar. Mapping high-dimensional data (like images) into low-dimensional representations (embeddings) has become a fairly common practice in machine learning these days.

**Introduction:** Face detection is a massively significant field in Computer vision and it is necessary task for applications like face recognition, face tracking, video observation, expression analysis and numerous other different fields. Given a discretionary picture, the objective of face identification is to decide if there are any countenances in the picture, and if present, return the picture area and degree of each face. The new issue of face location is the way to improve the location execution in unhindered situations. Since identifying faces in genuine, multi scale, dense scenes, small faces and blur pictures has numerous troubles including impediment, critical scale variety, distinctive light conditions, different facial postures, rich looks. To understand the project, we must understand few concepts used in the project such as computer vision, open-cv, image processing and image tracking.

A facial recognition is capable of matching a human face from an image or a video frame against a database of faces. Initially a form of computer application facial recognition system have many uses in recent times in the form of Technology such as Robotics and in the form of smartphones. While humans can recognize faces without effort, facial recognition is a challenging pattern recognition problem in computing. Facial recognition system attempts to identify a human face, which is three dimensional and changes in appearance of light and face expression, based on its two dimensional image. To this computational task face recognition system has four tasks.

1. Face detection is used to segment the face from the image background.
2. Segmented face is aligned to account of face pose, image size, Photographic properties
3. Feature extraction such as eyes, nose, mouth are pinpointed and measured in the image of database
4. Matching against a database of faces.

## FACE DETECTION:

Face recognition is perhaps the most utilized picture processing applications in the world. Actually following the profound learning philosophy, the machine is first shown the particular highlights of human appearances. Elucidating highlights, like the distance between the two eyes, the state of the normal human face, fill in as measurements to frame the face shape. In the wake of showing the human-explicit models of the face, it acknowledges all items in the picture that take after a similar shape as the face. The discovery of the face is made by making the particular measurements that make up the face human. Face discovery is a fundamental instrument for following clients in the shopping venture. After the face discovery measure, clients can be gathered to certain bunches to characterize their details. By this interaction, the quantity of clients and their principle highlights can be known to decide the most productive approach to expand deals.

Face detection algorithm is divided into steps:

A) Pre-Processing

B) Classification

C) Localization

# REVIEW OF FACE RECOGNITION METHODS

Face recognition methods divided into categories:

- Knowledge-based methods
- Feature-invariant methods
- Template matching methods

## Knowledge-based methods:

Knowledge-based methods are encoding our knowledge of human faces. These are rule-based methods. They try to capture our knowledge of faces, and translate them into a set of rules.

Its easy to guess some simple rules. For example, a face usually has two symmetric eyes, and the eye area is darker than the cheeks. Facial features could be the distance between eyes or the color intensity difference between the eye area and the lower zone. The big problem with these methods is the difficulty in building an appropriate set of rules. There could be many false positives if the rules were too general. On the other hand, there could be many false negatives if the rules were too detailed. A solution is to build hierarchical knowledgebased methods to overcome these problems. These methods show themselves efficient with simple inputs. But, what happens if a man is wearing glasses? There are other features that can deal with that problem. For example, there are algorithms that detect face-like textures or the color of human skin.

## Feature-invariant methods:

Feature-invariant methods that try o find invariant features of a face despite its angle or position. Facial recognition utilizes distinctives features of the face – including: distinct micro elements like: Mouth, Nose, Eye, Cheekbones, Chin, Lips, Forehead, Ears, Upper outlines of the eye sockets, the areas surrounding the cheekbones, the sides of the mouth, and the location of the nose and eyes. The distance between the eyes, the length of the nose and the angle of the jaw.

## Template matching methods

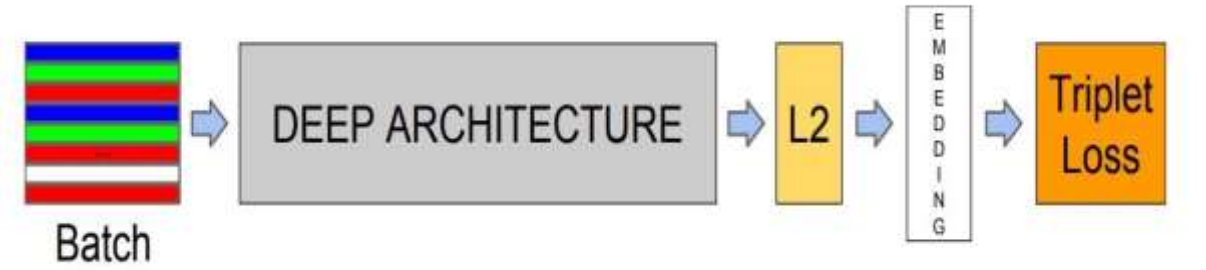
These algorithms compare input images with sorted pattern of faces or features. Template matching methods try to define a face as a function. One can try to find a standard template of all the faces. Different features can be defined independently. For example, a face can be divided into eyes, face contour, nose and mouth. Also a face model can be built by edges, But these methods are limited to facesthat are frontal. A face can also be represented as a shape.

## FaceNet :

FaceNet is the name of the facial recognition system that was proposed by Google researchers in 2015 in the paper titled FaceNet. A unified Embedding for face recognition and clustering. It achieved state of the art results in the many benchmark face recognition dataset such as Labeled faces in the Wild (LFW) and YouTube Face Database).

They proposed an approach in which it generates a high-quality face mapping from the images using deep learning architectures such as ZF-Net and Inception. Then it used a method called triplet loss as a loss function to train this architecture. Let's look at the architecture in more detail.

### Architecture :




*FaceNet Architecture*

FaceNet employs end to end learning in its architecture. It uses ZF-Net or Inception as its underlying architecture. It also adds several 1*1 convolutions to decrease the number of parameters. These deep learning models outputs an embedding of the image f(x) with L2normilization performed on it. These embedding then passed in to the loss function to calculate the loss. The goal of this loss function is to make the squared distance between two image embedding is independent of image condition and pose of the same identity is small, whereas the squared distance between two images of different is large. Therefore a new loss function called Triplet loss is used. This idea of using triplet loss in our architecture is that it makes the model to enforce a margin between faces of different identities.

## Triplet Loss :

The embedding of an images is represented by f(x) such as x\in \mathbb {R}. This embedding is in the form of vector of size 128 and it is normalized such that

$$\|f(x)\|_2^2 = 1$$

We want to make sure that the anchor image (x_{i}^a) of a person is closer to a positive image (x_{i}^p) (image of the same person) as compared to a negative image(x_{i}^n) (image of another person) such that :

$$\|f(x_i^a) - f(x_i^p)\|_2^2 + \alpha < \|f(x_i^a) - f(x_i^n)\|_2^2$$

$$\forall (f(x_i^a), f(x_i^p), f(x_i^n)) \in \top$$

Where \alpha is the margin that is enforced to differentiate between positive and negative pairs and T are the image space.

Therefore the loss function is defined as the following :

$$L = \sum_i^N \left[\|f(x_i^a) - f(x_i^p)\|_2^2 - \|f(x_i^a) - f(x_i^n)\|_2^2 + \alpha\right]$$

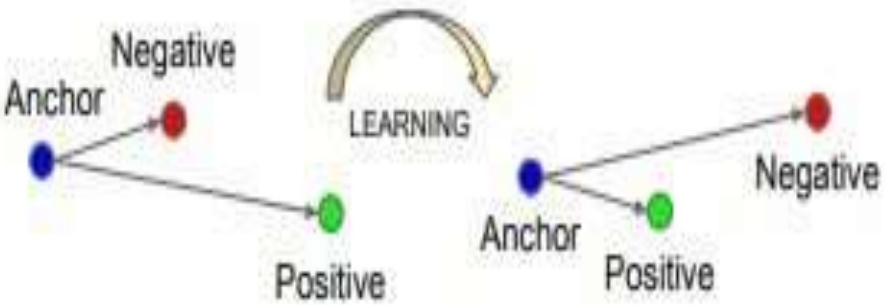
If triplets are easily satisfied above property then it would not helping training, so it is important to have the triplets that violate above equation.

## Triplet Selection :

In order to ensure faster learning, we need to take triplets that violate the equation above. This means for given x i^a we need to select triplets such that ||f(x{i}^{a})-f(x{i}^{p})||{2}^{2} is maximum ||f(x{i}^{a})-f(x{i}^{n})||{2}^{2} is minimum. It is computationally expensive to generate triplets based on whole training set. There are two methods of generating triplets.

Generating triplets on every step on the basis of previous checkpoints and compute minimum and maximum on a subset of data.

Selecting hard positive (x i^p) and head negative (x i^n) by using minimum and maximum on a mini batch



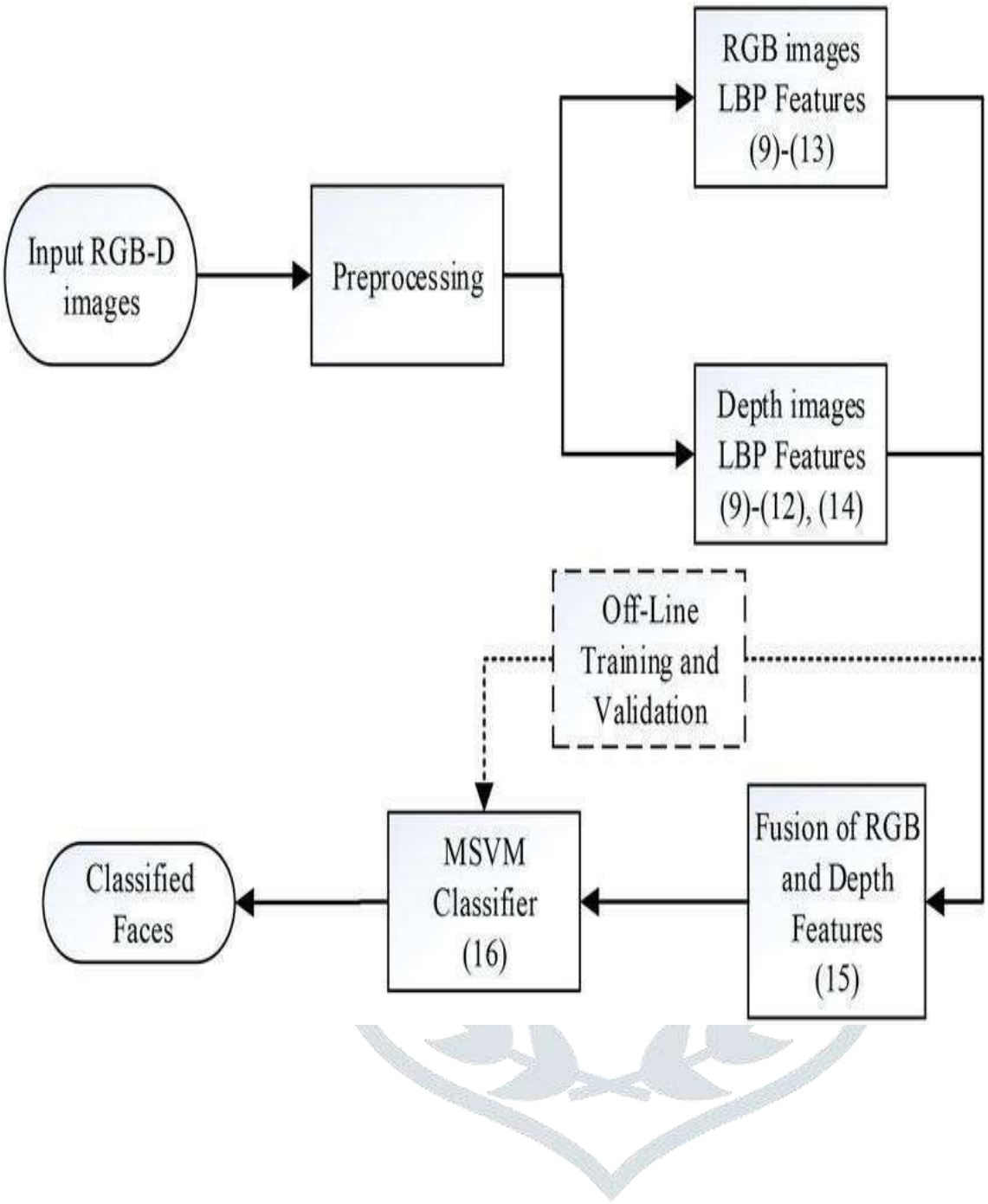

## Training:

This model is trained using Stochastic Gradient Descent (SGD) with backpropagation and Adagrad.

This model is trained on a CPU cluster for 1k-2k hours. The steady decrease in loss (and increase in accuracy) was observed after 500 hours of training. His model is trained using two networks :

- ZF-Net
- Inception

## Flow Chart of FaceNet




## Results :

This model uses 4 different types of architecture on Labeled faces in the wild and Youtube face DB dataset. These

- Labeled Faces in the Wild Dataset : This architecture uses standard, nonrestructed protocol on LFW dataset. First, this model uses 9 training splits to set L2 distance threshold value and then on the tenth split, it classifies the two images as same or different. There are two methods of preprocessing of the images out dataset on which accuracy is reported :
- Fixed Center crop of the image provided in LFW

- A face detector is used on LFW images if that fails then LFW face alignment is used. This model achieves a classification accuracy on 98.87% accuracy with 0.15% standard error and in the second case 99.63% accuracy with 0.09% standard error. This reduces the error rate reported by Deep Face by a factor of more than 7 and other state-of-threat Deep Id by 30%
- Youtube Face database : On youtube face dataset it reported an accuracy of 95.12% with standard error 0.39 using first 100 frames. It is better than 91.4% accuracy proposed by Deep Face and 93.5% reported by Deep Id on 100 frames.

## References :


1. M. Pantic and A. Vinciarelli, "Implicit Human-Centered Tagging," IEEE Signal Processing Magazine, vol. 26, no. 6, pp. 173-180, Nov. 2009.
2. Z. Zeng, M. Pantic, G.I. Roisman, and T.S. Huang, "A Survey of Affect Recognition Methods: Audio, Visual, and Spontaneous Expressions," IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 31, no. 1, pp. 39-58, Mar. 2009.
3. I. Laptev, M. Marszalek, C. Schmid, and B. Rozenfeld, "Learning Realistic Human Actions from Movies," Proc. IEEE Conf. Computer Vision and Pattern Recognition, pp. 1-8, June 2008.
4. J.A. Healey and R.W. Picard, "Detecting Stress during Real-World Driving Tasks Using Physiological Sensors," IEEE Trans. Intelligent Transportation Systems, vol. 6, no. 2, pp. 156166, June 2005.
5. M. Pantic, M. Valstar, R. Rademaker, and L. Maat, "Web-Based Database for Facial Expression Analysis," Proc. IEEE Int'l Conf. Multimedia and Expo, pp. 317-321, 2005.
6. E. Douglas-Cowie, R. Cowie, I. Sneddon, C. Cox, O. Lowry, M.McRorie, J.-C. Martin, L. Devillers, S. Abrilian, A. Batliner, N.Amir, and K. Karpouzis, "The HUMAINE Database: Addressing the

Collection and Annotation of Naturalistic and Induced Emotional Data," Proc. Second Int'l Conf. Affective Computing and Intelligent Interaction, A. Paiva et al., pp. 488500, 2007.

7. M. Grimm, K. Kroschel, and S. Narayanan, "The Vera am Mittag German Audio-Visual Emotional Speech Database," Proc. IEEE Int'l Conf. Multimedia and Expo, pp. 865-868, Apr. 2008.
8. G. McKeown, M.F. Valstar, R. Cowie, and M. Pantic, "The SEMAINE Corpus of Emotionally Coloured Character Interactions," Proc. IEEE Int'l Conf. Multimedia and Expo, pp. 1079-1084, July 2010.
9. S. Koelstra, C. Mu¨hl, M. Soleymani, A. Yazdani, J.-S. Lee, T. Ebrahimi, T. Pun, A. Nijholt, and I. Patras, "DEAP: A Database for Emotion Analysis Using Physiological Signals," IEEE Trans. Affective Computing, vol. 3, no. 1, pp. 18-31, Jan.-Mar. 2012.
10. M.F. Valstar and M. Pantic, "Induced Disgust, Happiness and Surprise: An Addition to the MMI Facial Expression Database," Proc. Int'l Conf. Language Resources and Evaluation, Workshop EMOTION, pp. 65-70, May 2010.